CLAIMS

What is claimed is:

1. A method for distributing data items from a particular set of data into a plurality of buckets based on distribution keys associated with said data items, the method comprising the steps of:

randomly selecting data items from said particular set of data to produce a sampled set of data items;

determining a plurality of ranges based on the distribution keys associated with the sampled set of data items;

assigning said plurality of ranges to said plurality of buckets; and

distributing each data item in said particular set of data to the bucket that has been assigned the range into which falls the distribution key of the data item.

2. The method of Claim 1 wherein the step of randomly selecting data items from said particular set of data includes randomly selecting data items from each subset of a plurality of subsets of said particular set of data.

3. The method of Claim 2 wherein the step of randomly selecting data items from each subset of a plurality of subsets of said particular set of data includes randomly selecting data items from each partition of a partitioned table.

4. The method of Claim 2 wherein the step of randomly selecting data items from each subset of a plurality of subsets of said particular set of data includes randomly selecting data items from subsets of data, stored in buffers in volatile memory, that represent results of one or more previously performed operations.

Sub A3

5. The method of Claim 1 further comprising the steps of:
assigning the plurality of buckets to a plurality of processes; and
causing each process of said plurality of processes to perform, in parallel with the other processes of said plurality of processes, an operation on the data items contained in any buckets assigned to the process.

6. The method of Claim 2 further comprising the step of selecting a distinct random seed for each subset of the plurality of subsets of said particular set of data.

7. The method of Claim 1 wherein:
the particular set of data is durably stored on a plurality of durable storage units; and
the step of randomly selecting data items from said particular set of data to produce a sampled set of data items includes randomly selecting durable storage units from said plurality of durable storage units and using the data items stored on said randomly selected durable storage units as the sampled set of data items.

8. The method of Claim 1 wherein the step of randomly selecting data items includes selecting a specified percentage of data items in said particular set of data.

9. The method of Claim 7 wherein the step of randomly selecting data items includes selecting a specified percentage of the plurality of durable storage units that are storing said particular set of data.

10. The method of Claim 8 further comprising the step of receiving, from a user, data that specifies said percentage.

11. The method of Claim 9 further comprising the step of receiving, from a user, data that specifies said percentage.

12. The method of Claim 5 wherein said operation is specified in a database command, the method further comprising receiving with said database command data that indicates how much of said particular set of data to randomly select to produce said sampled set of data items.

13. The method of Claim 1 wherein the step of determining a plurality of ranges based on the distribution keys associated with the sampled set of data items includes determining ranges that contain an approximately equal amount of distribution keys associated with said sampled set of data items.

14. A computer-readable medium carrying instructions for distributing data items from a particular set of data into a plurality of buckets based on distribution keys associated with said data items, the instructions comprising instructions for performing the steps of:

randomly selecting data items from said particular set of data to produce a sampled set of data items;

determining a plurality of ranges based on the distribution keys associated with the sampled set of data items;

assigning said plurality of ranges to said plurality of buckets; and

distributing each data item in said particular set of data to the bucket that has been assigned the range into which falls the distribution key of the data item.

Sub A3

15. The computer-readable medium of Claim 14 wherein the step of randomly selecting data items from said particular set of data includes randomly selecting data items from each subset of a plurality of subsets of said particular set of data.

16. The computer-readable medium of Claim 15 wherein the step of randomly selecting data items from each subset of a plurality of subsets of said particular set of data includes randomly selecting data items from each partition of a partitioned table.

17. The computer-readable medium of Claim 15 wherein the step of randomly selecting data items from each subset of a plurality of subsets of said particular set of data includes randomly selecting data items from subsets of data, stored in buffers in volatile memory, that represent results of one or more previously performed operations.

18. The computer-readable medium of Claim 14 further comprising instructions for performing the steps of:

assigning the plurality of buckets to a plurality of processes; and

causing each process of said plurality of processes to perform, in parallel with the other processes of said plurality of processes, an operation on the data items contained in any buckets assigned to the process.

19. The computer-readable medium of Claim 15 further comprising instructions for performing the step of selecting a distinct random seed for each subset of the plurality of subsets of said particular set of data.

20. The computer-readable medium of Claim 14 wherein:

the particular set of data is durably stored on a plurality of durable storage units; and

the step of randomly selecting data items from said particular set of data to produce a sampled set of data items includes randomly selecting durable storage units from said plurality of durable storage units and using the data items stored on said randomly selected durable storage units as the sampled set of data items.

21. The computer-readable medium of Claim 14 wherein the step of randomly selecting data items includes selecting a specified percentage of data items in said particular set of data.

22. The computer-readable medium of Claim 20 wherein the step of randomly selecting data items includes selecting a specified percentage of the plurality of durable storage units that are storing said particular set of data.

23. The computer-readable medium of Claim 21 further comprising instructions for performing the step of receiving, from a user, data that specifies said percentage.

24. The computer-readable medium of Claim 22 further comprising instructions for performing the step of receiving, from a user, data that specifies said percentage.

Sub A3

25. The computer-readable medium of Claim 18 wherein said operation is specified in a database command, the computer-readable medium further comprising instructions for receiving with said database command data that indicates how much of said particular set of data to randomly select to produce said sampled set of data items.

26. The computer-readable medium of Claim 14 wherein the step of determining a plurality of ranges based on the distribution keys associated with the sampled set of data items includes determining ranges that contain an approximately equal amount of distribution keys associated with said sampled set of data items.